How-To: Run a Load for a Customer Using the Baldur Pipeline (old version)

> This guide assumes you are already familiar with how to login to one of the EDH terminal servers and are generally familiar with the EDH CLI commands. For further information please see Using the EDH CLI
>
> The examples below are from a single load of Northwell in the production environment. Please alter these commands for the customer, environment, and profile as necessary. See Customer for a list of the valid customers.
>
> It is advised that for jobs expected to run for longer than 10-15 minutes, you should initialize them in a tmux shell. This will prevent the job from failing if your internet connection is interrupted. More info at A Gentle Introduction to tmux
>
> By default all jobs run in PROD will report a success or failure message to the #edh-job-stats channel in slack. You may also add the -S argument to any of the commands below to force that behavior in STAGE or QA.
>
> You can also type "edh --help" for more information on the commands shown in these examples.

Experian

Experian jobs that process the whole file do not require a customer to be specified. All other jobs in the pipeline require a customer parameter.

Step 1) Cleanse the latest Experian quarterly file

(You may skip this step if the most recent Experian file has already been cleansed)

```
edh experian cleanse -e prod -p xlarge
```

By default the job will look at

`/media/edh1/experian/input/`

to find the incoming experian data.

Step 2) Run the cleansed Experian data through Anchor NCOA and CASS

(You may skip this step if the most recent Experian file has already been run through anchor)

```
edh experian anchor -e prod -i "/media/edh1/experian/output/*.txt"
```

This is not a spark job so it does not need a spark job profile to be specified.

Step 3) Load the raw experian information for the customer's service areas into Cassandra

```
edh experian raw -e prod -p large -c northwell
```

By default this job will look in

`/media/edh1/baldur/output`

to find the post-anchor experian files. They should be prefixed by "experian" and be in their unzipped .txt format.

Step 4) Create a prospect build for the customer using the latest Experian data

```
edh experian prospect -e prod -p large -c northwell
```

Utilization

Step 5) Cleanse the latest utilization data for the customer

First ensure that the desired utilization data has been copied into:

`/media/edh1/northwell/utilization`

Then execute the following command

```
edh load cleanse -e prod -p large -c northwell -C northwell.config
```

Step 6) Run the cleansed utilization data through Anchor NCOA and CASS

```
edh load anchor -e prod -c northwell -i
"/media/edh1/northwell/output/baldur_ToAnchor_20001_NORTHWELL
HEALTH~001464_20180215T190812.778Z.txt"
```

Step 7) Load the cleansed utilization data with anchor enrichments into Cassandra

> Given the current design of the system, input files of more than 5 million rows tend to perform poorly and occasionally cause the load identify job to fail. For this reason it is recommended that larger input files are split into files containing 5 million rows or less.

To do this, first move the anchor output file into a working directory and unzip it

```
mv "/media/edh1/baldur/anchor_download/output/baldur_NORTHWELL
HEALTH~001464_output_20180215T190812778Z.zip" /media/edh1/northwell/input/
cd /media/edh1/northwell/input
unzip -d identify baldur_NORTHWELL\
HEALTH~001464_output_20180215T190812778Z.zip
  Archive: baldur_NORTHWELL HEALTH~001464_output_20180215T190812778Z.zip
  inflating: identify/baldur_20180215T190812778Z.txt
cd identify
wc -l baldur_20180215T190812778Z.txt
  27645883 baldur_20180215T190812778Z.txt
split -d -l 5000000 baldur_20180215T190812778Z.txt northwell-split
```

Please note that the above steps are very slow on the current shared filesystem. This process took close to 1 hour to unzip and split a 22GB file containing 27645883 lines. Inside the working directory you should have the following:

```
ls -alh
  total 43G
  drwxr-xr-x  2 root root 4.0K Feb 27  2018 .
  drwxr-xr-x 10 root root 4.0K Feb 27 20:36 ..
  -rw-r--r--  1 root root  22G Feb 15 20:26 baldur_20180215T190812778Z.txt
  -rw-r--r--  1 root root 3.9G Feb 27 22:15 northwell-split00
  -rw-r--r--  1 root root 3.9G Feb 27  2018 northwell-split01
  -rw-r--r--  1 root root 3.9G Feb 27  2018 northwell-split02
  -rw-r--r--  1 root root 3.9G Feb 27  2018 northwell-split03
  -rw-r--r--  1 root root 3.9G Feb 27  2018 northwell-split04
  -rw-r--r--  1 root root 2.1G Feb 27  2018 northwell-split05
```

Once you have your input files prepared, run the command below for each file

```
edh load identify -e prod -p xlarge -c northwell -i
/media/edh1/northwell/input/identify/northwell-split00
```

Enrichments

Step 8) Enrich person records with residence information (assign household preferences)

```
edh enrich residence -e prod -p large -c northwell
```

Based on the way the current job is implemented, this step must be run twice to assign household_ids correctly (this will be fixed in an upcoming refactor)

```
edh enrich residence -e prod -p large -c northwell
```

Step 9) Enrich person records based on presence of children in household

```
edh enrich children -e prod -p large -c northwell
```

Step 10) Enrich person records with customer location preferences

```
edh enrich locations -e prod -p large -c northwell
```

Step 11) Enrich persons with propensity model scores

```
edh enrich propensities -e prod -p large -c northwell
```

Step 12) Enrich persons with recipe scores

```
edh enrich recipes -e prod -p large -c northwell
```

Collapse Duplicate Persons

Step 13) Collapse multiple different persons from the same load

(You only need to run this if multiple sources that could contain the same persons were loaded within the same batch, otherwise the identify step should cover this functionality)

```
edh collapse persons -e prod -p large -c northwell
```

Health Check

Step 14) Check the health of the data for the current customer

```
edh check health -e prod -p large -c northwell
```

Updating Redshift and Elasticsearch

Step 15) Push the latest copy of all cutomer data to Elasticsearch

```
edh refresh elasticsearch -e prod -p large -c northwell --num-shards 5
--num-replicas 1
```

Step 16) Push the latest copy of all customer data to Redshift

```
edh refresh redshift -e prod -p large -c northwell
```